Process of Developing Predictive Models

Roosevelt C. Mosley, Jr.
Emily Stoll
PROCESS OF DEVELOPING PREDICTIVE MODELS
Process of Developing Predictive Models

- Strategy
- Data Preparation
- Predictive Modeling Techniques
- Building a Generalized Linear Model (GLM)
- Implementation
Predictive Analytics Strategy - Considerations

- Goals
- Application
- Implementation
- Approach
- Software
- Data
Goals

- **Question:** What am I trying to accomplish by using predictive analytics?
- **Potential answers**
  - **Project driven** (I want to develop a new rating plan, I need to detect claim leakage)
  - **Externally driven** (I see a lot written and hear a lot of presentations on predictive analytics)
  - **Defensive** (avoid anti-selection, my competitors are doing it)
  - **Goal driven** (We want to become a data and analytics driven company, we want to use analytics to support our vision – value, enhancing financial security, we want to use analytics to create a competitive advantage)
Predictive Models – One Tool in the Toolkit
Data Preparation
Data is Increasing Exponentially
Challenges Associated with the Expanding Data Universe

- Data historically collected for a number of different purposes – except analytics
- Creates challenges
  - Missing information
  - Incorrect data
  - Significant time required to process data
- Intentional data processing
  - Identify the right data
  - Collect and store data consistently and accurately
  - Prepare data once for multiple applications
Missing/Unknown Values

- **Causes**
  - Data collection error/issue
  - External data joins
  - Rating plan design

- **Potential remedies**
  - Collect data
  - Include with a valid level
  - Model as a separate level
  - Imputed values
# Accurate Exposure Information

<table>
<thead>
<tr>
<th>Policy Number</th>
<th>Age</th>
<th>Territory</th>
<th>Symbol</th>
<th>Effective Date</th>
<th>Expiration Date</th>
<th>Transaction Date</th>
<th>Exposure</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>45</td>
<td>10</td>
<td>16</td>
<td>1/1/2000</td>
<td>7/1/2000</td>
<td>1/1/2000</td>
<td>0.50</td>
</tr>
<tr>
<td>1</td>
<td>45</td>
<td>10</td>
<td>16</td>
<td>1/1/2000</td>
<td>7/1/2000</td>
<td>4/1/2000</td>
<td>-0.25</td>
</tr>
<tr>
<td>1</td>
<td>45</td>
<td>10</td>
<td>19</td>
<td>1/1/2000</td>
<td>7/1/2000</td>
<td>4/1/2000</td>
<td>0.25</td>
</tr>
</tbody>
</table>

Claim on 5/1/2000

![Diagram](image)
Training/Testing/Validation Data

- Develop model with training data
- Test the model developed using testing data
  - Assists in developing optimal model
- Validate final model using validation data
  - Helps decide between competing models
  - Ensures final model is appropriate
- Split percentages
  - Depends on dataset size
- Training and test split only is also acceptable standard practice
Basic Data Summarizations

- Processes used
  - Distribution graphs
  - Simple location measures
  - Two way graphs
  - Data summary tables

- Helps with
  - Understanding
  - Cleaning
  - Potential applications
Predictive Modeling Techniques

- Generalized linear models
- Decision trees
- Gradient boosting
- Neural networks
- Machine Learning
- Ensemble
Machine Learning Overview

**Supervised**
- Predictive
- Target Variable
- Task Driven
- Regression, Classification

**Unsupervised**
- Descriptive
- No Target Variable
- Data Driven
- Clustering, Pattern Discovery, Dimension Reduction

Reinforcement – Algorithm Learns to React
Choice of Models

- Underlying data
- Purpose of the analysis
  - Target variable structure
  - Use of the results
  - Prediction vs. understanding
- Sophistication of user
- Trial and Error
  - Run multiple models and evaluate fit
PROCESS OF DEVELOPING PREDICTIVE MODELS
Remember the Goal

- **Find the signal, remove the noise**

  \[
  \text{Response Variable} = \text{Systematic Component} + \text{Random Component}
  \]

  **Signal**

  **Noise**

  - **Underfit**
  - **Overfit**

- **By balancing over- and under-fitting**
Model output also includes statistics that help a modeler make modeling decisions.
Control Variables and Offsets

- Some variables may need to be controlled for in the model
  - Time
  - Geography

- A modeler may want to include known effects in the model as offsets
  - Coverage type variables: limits, deductibles, AOIs
  - Discounts marketed or required to be a certain level
  - Portions of a rating plan not being analyzed

\[
E[Y_i] = g^{-1}(\sum X_{ij} \beta_i + \xi_i)
\]
“It is not generally possible to determine from a single GLM which set of factors are significant since the inclusion or exclusion of one factor will change the observed effects and therefore possibly the significance of other correlated factors in the model. To determine the theoretically optimal set of factors, therefore, it is generally necessary to consider an iterated series of models.”

Distributional Bias

Youngest Age on Policy by Number of Operators on the Policy (Cramer’s V = 0.397)

Exposures scaled to 100% show significant distributional bias

Youngest Age on Policy by Collision Deductible (Cramer’s V = 0.050)

Exposures scaled to 100% show almost no distributional bias

23 Exposures scaled to 100% show significant distributional bias
Modeling Effect of Distributional Bias

Youngest Age on the Policy
Blue line: Number of operators excluded from model
Green line: Number of operators included in model

The addition of a correlated variable changes the parameters of the age curve.

The addition of a less correlated variable barely changes the parameters of the age curve.
Modelers decide which predictor variables to include, and how to include them.

Statistics are helpful in making decisions, but are not always definitive.
Different modelers may use different approaches or make different modeling decisions on the same data, and both build reasonable models.
A predictor variable should only be included in the model if the variable has a \textit{systematic effect} on the response variable.

- **How can a modeler tell?**
  - Standard errors
  - Type III tests
  - Consistency over time
  - Use of a test dataset
Parameters and Standard Errors

- **Simplification:**
  - Horizontal line test can inform what levels to group

- **Inclusion/Exclusion:**
  - If all relativities are close to 1.0 and/or standard errors are wide, it indicates exclusion of a variable

Using only standard errors to make modeling decisions will tend to lead to an under fit model
χ² Test

- Use chi test to compare nested models
  - H₀: Two models are basically the same
  - H₁: Two models are statistically different
- If models are the same, use simpler model
- Low χ² (~ < 5%): reject H₀

Testing for Inclusion:
Model 1: Base Model
Model 2: Base Model + Var Y
  Low χ² indicates the models are not the same and Var Y is statistically significant.

Testing a Simplification:
Model 1: Base Model w/ Var Y
Model 2: Base Model w/ Simplified Var Y
  High χ² indicates the models are basically the same, so the simplification has not removed significant signal.

Using only χ² to make decisions will tend to lead to an over fit model
Consistency

- **Inclusion/Exclusion:**
  - Looking for similar pattern over different years or subsets of data

- **Simplification:**
  - Create groups or fit curves to capture consistent trend
Using a Test Dataset

- **Balance test**
  - Compare actual values to predicted values on a holdout dataset to find areas of systematic imbalance

- **Stability test**
  - Refit model structure on holdout dataset to compare parameters

Use these results to go back and retest specific variables or levels of variables.
Model Validation

- Score a hold-out dataset with final model parameters
- Actual and predicted values should line up in aggregate

Which model is better at differentiating between high and low risks?

Do high fitted values correspond to high actual values?

Models can be compared to see which is more predictive, but predictive power is not the only consideration when implementing a model.
# Model Implementation

<table>
<thead>
<tr>
<th>Considerations</th>
<th>Potential Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regulations</td>
<td></td>
</tr>
<tr>
<td>ASOPs</td>
<td>State specific model refit</td>
</tr>
<tr>
<td>Cost of implementation</td>
<td></td>
</tr>
<tr>
<td>Customer impact</td>
<td>Modified countrywide model refit</td>
</tr>
<tr>
<td>Competitive framework</td>
<td></td>
</tr>
<tr>
<td>Explainability</td>
<td>Selections off of indicated factors</td>
</tr>
<tr>
<td>Credibility</td>
<td></td>
</tr>
</tbody>
</table>
Questions
For More Information

For more information, contact Marc Rosenberg, senior casualty policy analyst,
At rosenberg@actuary.org or (202) 785-7865